基于商品描述文案的点击预测模型 *

黄皓炫，盛 武†

(安徽理工大学 经济与管理学院，安徽 淮南 232000)

摘 要：为了预测商品描述文案中商品特征对点击的影响，量化分析用户的消费行为特征及缓解冷启动问题，建立了一种基于 LDA 模型和文本情感分析的点击预测模型。该模型基于 LDA 主题模型对商品描述词的分类筛选，对构成词进行情感分析，构建特征向量以表示用户对商品各特征的情感倾向，并通过 LightGBM 算法进行对点击的预测。模型可以将非结构化文本数据转换为结构化数据，量化用户对商品不同特征的兴趣倾向，并利用不同商品的相似特征缓解冷启动问题。实验结果表明模型有效提高了点击预测效果并能缓解冷启动问题。

关键词：LightGBM；点击预测；文本情感分析；LDA 主题模型；冷启动；

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2022.01.0025

Click prediction model based on product description

Huang Haoxuan, Sheng Wu†

(*Dept. of Economics & Management*, *Anhui University Of Science & Technology*, *Huainan Anhui* 232000, *China*)

Abstract: In order to predict the impact of commodity characteristics on click in commodity description copy, quantitatively analyze users' consumption behavior characteristics and alleviate the cold start problem, this paper established a click prediction model based on LDA model and text emotion analysis. Based on the LDA topic model, the model classifies and screens the commodity description words, analyzes the emotion of the constituent words, constructs the feature vector to represent the user's emotional tendency to the characteristics of the commodity, and predicts the click through the lightgbm algorithm. The model can transform unstructured text data into structured data, quantify users' interest in different characteristics of goods, and use the similar characteristics of different goods to alleviate the cold start problem. The experimental results show that the model can effectively improve the click prediction effect and alleviate the cold start problem.

Key words: lightgbm; click prediction; text sentiment analysis; LDA topic model; cold start;

0 引言

网购平台上有很多商家等撰写的商品描述文案，其中往往包括了对商品的外观、尺寸、颜色、功能、打折信息等多方面多角度的详细描写，体现了行业从事者对消费者核心需求的思考、对同行及自己产品卖点特征的判断。产品的属性特征是多方面的，对消费者的吸引力也各有区别，有研究表明，商品描述的差异会影响消费者的购物意愿[1,2]。通过研究商品描述及其商品点击量的不同，可以了解到消费者对商品不同属性点的偏好和需求的差异。相比能够直接体现用户对商品主观感受的购物后的用户评价，购物前的商品描述更能反映消费者的消费冲动，体现了消费者的核心需求。对商品描述的研究，不仅可以为消费者更高效获取商品信息提供支持，也能为商家改善商品性能、研发新产品、调整商品卖点宣传提供依据。

目前国内外专门针对电商领域的中文商品描述的研究比较少，与此研究问题涉及内容相近的研究主要有计算广告领域的点击率预测研究、推荐系统领域的评分预测研究等。点击率预测是计算广告领域中一个重要的研究内容[3]。因为按点击付费是互联网广告的主要计价模型之一，从而通过对点击率的预测研究，可以提高广告主的投资回报率的同时，最大化用户对展示广告的满意程度[4]。点击率预测模型主要分为基于历史日志的预估模型和基于稀疏数据的预估模型。前者基于广告的丰富的历史数据(如广告的位置、内容等)，然后通过逻辑回归、贝叶斯网络等算法从而进行对点击率的预测[5~7]。但这些方法的缺点是难以处理稀疏数据型的广告或新广告，因此诞生了后者如基于层次聚类分析、相似项、因子分解机等方法的预估模型[8-10]。文献[4]就从广告语义的角度出发，通过 LDA 主题模型以挖掘广告文本中的主题，以广告与主题的相关性基于 FM 模型建立了点击率预测模型，证实了文本语义和点击的相关性。

推荐系统是在大量数据中筛选出最符合用户需求偏好的结果给用户的一种系统[11, 12]。其中，协同推荐算法作为推荐系统中最主流的算法之一，主要通过用户对项目的评分来研究用户和项目之间的关联进行预测[13]。不过，早期的推荐系统算法主要将商家视为一个商品，通过寻找相似商品或相似用户进行推荐。而随着互联网的发展、社交网络的兴起，用户和商户的互动在不断增加，评论信息数量不断攀升，文献[14]通过分析用户的评论建立评分矩阵，提出了一种基于高斯模型的优化算法来研究用户在商品不同方面的偏好。而文献[15]从常用词或形容词的角度建立词袋的角度构建评分预测模型，文献[16]则通过 LDA 主题模型提取评论的主题特征分布作为自变量构建评分预测模型。这些方法根据对用户评论文本的分析处理，探究了文本信息与评分的关联性，从评论文本语义的角度构建了评分预测模型。针对如何进一步提高评分预测的精度，有学者通过融合其他因素或方法来解决这个问题，并获得了良好的效果。文献[17]提出结合融合元数据和评分数据构建特征变量进而进行对评分的预测。文献[18]

收稿日期：2022-01-12；**修回日期：**2022-03-08 **基金项目：**安徽省自然科学基金资助项目(1808085MG212)、安徽省高等学校省级教学示范课基金项目

作者简介：黄皓炫(1995-)，男，广东韶关人，硕士研究生，主要研究方向为大数据、数据分析(jmmark@vip.qq.com)；盛武(1969-)，男(通信作者)，安徽涡阳人，副教授，硕导，博士，主要研究方向为管理决策与预测、大数据、安全管理(wsheng@aust.edu.cn).

则基于文本情感分析，对文本数据进行情绪挖掘与分析，从而提取文本中的主要观点倾向，将其作为自变量构建了评分预测模型，并取得了较好的评分预测效果。

在以上相关研究中，本文与文献[16,18]的研究内容较为接近，都是通过对非结构化文本信息进行分析从而构建预测模型。其中主要区别如下：

a)预测的目标不同。文献[16,18]都是根据评论构建评分预测模型，关注的是商品的售后口碑，本文则通过对商品描述文案不同特征的情感分析及 LightGBM 的可解释性构建点击模型，更关注商品中不同属性对消费者的吸引力影响。

b)特征量化角度不同。文献[16]通过使用 LDA 主题模型对文档词语进行主题分类，以分词出现的概率作为各特征的量化值。本文考虑到商品不同功能对消费者的吸引力不同，以商品各特征的情感倾向作为量化值，进一步提高了预测效果。

c)特征值的量化不同。在情感分析中，情感词的确立及情感权重的加权都是十分重要的。文献[19,20]通过基于评论文本中通用情感词典的积极情感词、消极情感词等情感词进行对整段评论的情感分析。但对于商品描述文案而言，文本主要由对商品的描述词语组成，以功能性词和积极情感词为主，通用的情感词典无法反映消费者的情感倾向。本文通过对商品特征进行分解，以与商品特征关联度较高的词作为情感词，再通过定义一个情感倾向计算公式作为消费者的情感倾向权重，因而具有普适性，不需要预定的情感词典，并可应用到不同商品的不同特征。

d)冷启动问题。传统协同过滤算法利用用户对商品的评分数据做推荐，存在数据稀疏性和冷启动问题[21]。本文基于对商品描述文本的挖掘，以用户商品的特征的偏好构建预测模型，因此可以通过具有相似特征的商品的数据模型来解决新商品缺乏数据的问题，从而缓解商品的冷启动问题。

基于此，本文通过分析商品描述文本及点击量的之间的关系，提出一种基于商品描述文案的点击预测模型。本文先利用 jieba 分词对商品描述文本进行词语级分割，以及通过停用词去除无关词语；然后利用 LDA 主题模型提取商品隐含特征，建立商品的属性分类；再基于各词汇的概率分布及权重量化文本情感值，将商品描述文本特征量化；最后通过 LightGBM 算法模型对商品的点击进行分类预测，分析商品各项特征对点击量的影响，挖掘用户的行为特征，并缓冷启动问题。

1　基于商品描述的点击预测模型设计

本文基于 LightGBM 和文本情感分析的点击预测模型主要包括数据预处理、特征提取、文本情感分析、LightGBM 模型训练和结果分析五个部分，模型框架如图 1 所示。

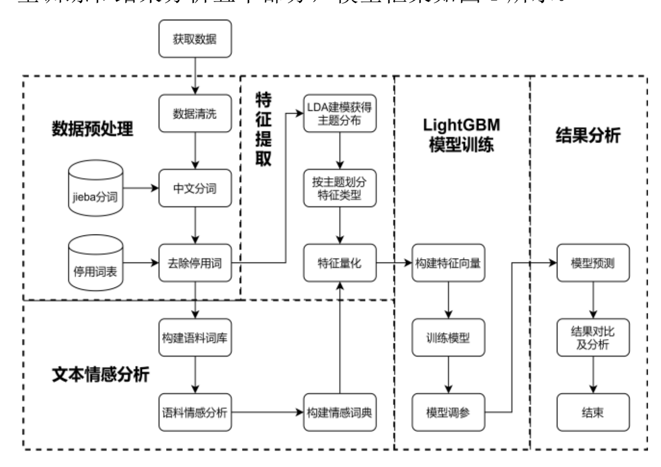


图 1　基于商品描述文案的点击预测模型

Fig. 1　Click prediction model based on product description

1.1　特征提取

不同于便于提取分析的结构化数据，商品描述文案的结构不规则，不符合预设的既定处理方法，属于非结构化数据。其中，在对原始语句进行中文分词和去除停用词后，本文通过 LDA 主题模型进行对词语的主题分类，从而获得研究目标的主题分类，以其作为目标的特征属性，再进行下一步分析。LDA 主题模型由 Blei、David M.、Ng,Andrew Y.等于 2003 年提出的，一种基于词袋模型的分析文档主题分布的一种三层贝叶斯概率模型[22]。它假设一篇文章具有 K 个主题，而每个主题又对应不同的词。因此文档的生成如下：

a)从狄利克雷分布 α 中取样生成文档 i 的主题分布 θ_i 。

b) 从主题的多项式分布 θ_i 中取样生成文档 i 的第 j 个的主题 $z_{i,j}$ 。

c) 从狄利克雷分布 β 中取样生成主题 $z_{i,j}$ 对应的词语分布 $\phi_{z\,i,j}$ 。

d) 从词语的多项式分布 $\phi_{z\,i,j}$ 中取样生成最终词语 $\omega_{i,j}$ 。

重复步骤 b)~d)从而生成文档 i。

基于此，LDA 主题模型通过逆向该过程，即给定文档 i 及词语，然后通过吉布斯采样(Gibbs sampling)方法反推其主题的分布。从而获得文档 i 的 K 个主题，及组成主题的词语组。根据文档划分的 K 个主题，商品 I_i 的特征词组可以记为 η_i ， $\eta_i=[\eta_{i1},\eta_{i2},\cdots,\eta_{iK}]$ 。其中， η_{iK} 表示商品 I_i 中与主题 K 的相关性词的组合，若不存在相关词，则 η_{iK} 为空集。其中，相关性词为与主题 K 相关性最高的前 1000 个词。模型结构如图 2 所示。

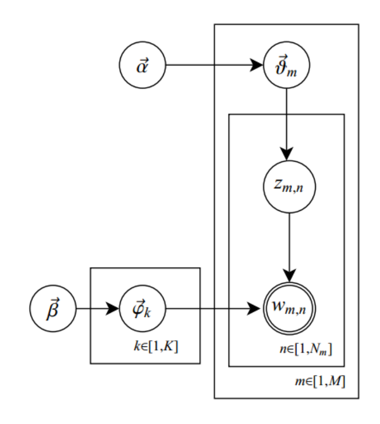


图 2　LDA 模型结构

Fig. 2　LDA topic model

1.2　文本情感分析及特征量化

文本情感分析，又称倾向性分析或意见挖掘，是通过计算、分析、归纳文本信息，从而获得其中的观点、情绪或倾向的过程。根据粒度细分的不同，可以分为篇章级、句子级和词语级三个层次，即对一篇文章、一个句子或一个词的情感倾向分析。本文在获得预处理后的文本后，把每个商品描述文案的点击量作为其对应的情感倾向，然后基于统计方法进行情感分析，从而获得词语的情感倾向。由于点击量分布的位置平均数靠左，峰度陡峭，数值范围跨度大，若直接取其词语的数学期望作为其情感倾向会导致高点击量的权重过大，因此取词语的点击量进行对数处理后的数学期望作为词语的情感倾向，点击量分布如图 3 所示。

最后，词语的情感倾向计算式(1)定义如下：

$$\omega_t=\frac{1}{V}\sum_{v=1}^{V}\log_{10}C_v \tag{1}$$

其中， ω_t 表示词语 t 的情感值，V 表示包含词语 t 的商品描述文案的频数， C_v 表示第 v 个词语的点击量。

根据商品 T_i 的特征词组 η_i ，通过情感倾向计算公式则可获得商品 I_i 的特征向量 ψ_i ， $\psi_i=[\psi_{i1},\psi_{i2},\ldots,\psi_{iK}]$ 。其中， ψ_{iK} 表

示商品 I_i 的第 K 个特征的特征值。特征值的计算公式(2)如下：

$$\psi_{iK}=\sum_{1}^{T}w_t \tag{2}$$

其中，T 表示相关词特征词组 η_{iK} 中的词语数。若特征词组 η_{iK} 中没有其特征词，则 $\psi_{iK}=0$。特征量化的过程如图 4 所示。

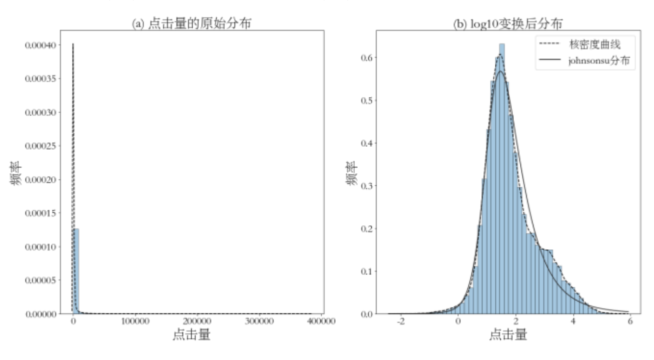


图 3　点击量分布图

Fig. 3　Click distribution

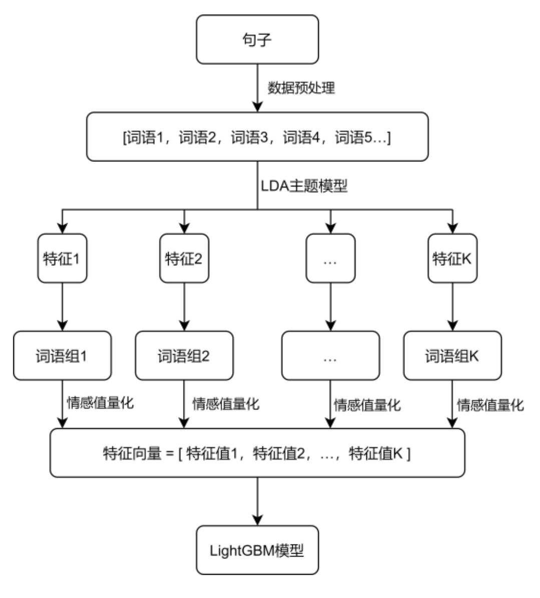


图 4　特征量化过程

Fig. 4　Feature quantization process

1.3　LightGBM 算法

在上述特征向量构建完成后，将特征向量作为自变量输入 LightGBM 模型。LightGBM 模型是由微软于 2017 年开源的一种基于决策树的集成算法[23]。相较于 XGBoost、GBDT 等算法在计算信息增益时需要扫描所有样本以找到最优划分点，LightGBM 模型采用了 Histogram、GOSS、EFB 等算法方法，从而在面对大量数据或者特征维度很高的数据集时，具有更快的训练速度、更低的内存消耗和更好的准确率等优点。

Histogram 算法，通过对每个特征进行分箱(bin)处理，构造成一个宽度为 k 的直方图，在遍历数据时，根据分箱在直方图中累积统计量，根据遍历后的累计统计量，遍历寻找最优的分割点。

GOSS 算法，又名单边梯度采样算法，从减少样本的角度出发，根据信息增益的定义，排除大部分对信息增益影响小的梯度小的样本，保留梯度大的样本。GOSS 算法会先将进行分裂的特征的所有取值按绝对值大小降序排序，选取绝对值最大的 a*100%个样本，再从剩余数据中随机选取 b*100%的样本，并乘以一个常数，从而减少改变原数据集分布的影响。

EFB 算法，可以通过将一些特征进行融合绑定，从而降低特征数量。LightGBM 的 EFB 算法将独立特征绑定转换为图着色问题，构建一个加权无向图，将所有特征视为图的各个定点，将不相互独立的特征用一条边链接，边的权重即为两个相互连接的特征的总冲突值，选取冲突小的特征融合，从而解决数据稀疏问题。

2　实验过程

2.1　数据获取

本文实验数据取自阿里云天池实验室的公开数据“Product Description”[24]，以其中 content 数据包中商品的描述文案及各类用户的点击量信息。本文所用点击量若无特别说明均为所有类型用户点击量之和，且缺失值视为 0。其中，筛选以“外套”为关键词的数据作为研究对象，共包含 34892 条数据，其点击量分布如图 3 所示，其点击量百分位数如表 1 所示。

表 1　点击量百分位数

Tab. 1　Click percentile

百分位	0%	25%	50%	75%	离群值线	100%
点击量	0.01	17.38	45.36	210.59	500.41	382745.58

根据样本的点击量分布，将点击量划分为 2 类：普通点击量，高点击量。由于商品的推荐算法大多以用户的点击率(CTR)为主要优化目标，这导致当某一类商品的点击率越高时，就会得到更多的曝光，因此点击量往往呈现两极分化的趋势。一般而言，曝光度低的商品的点击量主要与其商品的属性相关，而曝光度高的商品的点击量则容易受到各方面的影响。由此，本文对点击量的划分以离群值的判定为基础。离群值，也称溢出值，一般是指数据中与其他观察值具有明显不同特征的那些观察值。在此以四分位法划分离群值线，其计算式(3)如下：

$$Outl=Q_{75}+(Q_{75}-Q_{25})\times 1.5 \tag{3}$$

其中，Q_{25}，Q_{75} 分别代表样本中点击量从小到大排列后的第 25%和第 75%的值。记普通点击量为 0，高点击量为 1，则普通点击量的样本数为 28667，高点击量的样本数为 6225，比值为 4.61，因此本样本为不平衡样本。

2.2　数据处理及特征量化

本文在获取商品描述文本后，数据处理流程如下：

文本预处理。对商品描述文本进行预处理，对数据进行清洗，筛选出目标数据集；然后对商品描述文本信息进行 jieba 分词，将每个样本的文本转换为词组；再通过停用词表过滤掉不重要的词语，如“的”“啊”“！”等助词和符号。

特征提取。对预处理完的文本建立词典，使用 LDA 主题模型进行主题分析。对不同主题数进行分别迭代对比，其中，当主题数 num_topics=6 时，主题比较清晰，主题分布如图 5 和图 6 所示。



图 5　主题词云图

Fig. 5　Theme word cloud

如图 5 主题词云图所示，6 个主题可大致标记为“风格”、“设计”“保暖”“换季”“穿搭”“身材”6 个特征。而在主题分布图中，圆圈表示不同的主题以及它们之间的距离，类似的主题看起来更近，而不同的主题更远，图中主题圆的相对大小对应于语料库中主题的相对频率。如图 6 主题分布图所示，6 个主题的气泡较为分散，仅主题 1 和主题 3 有少量相交部分，证明该主题划分较为清晰独立，有较高的区分度。

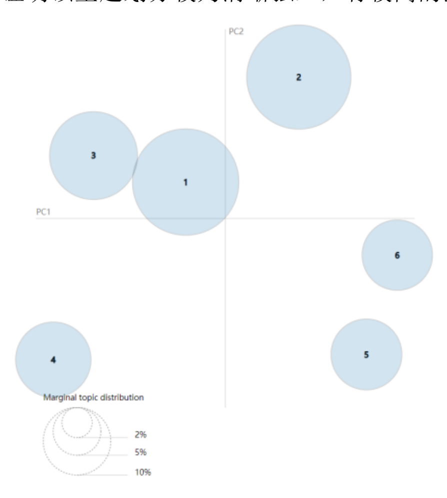


图 6　主题词分布

Fig. 6　Topic distance map

特化。根据预处理后的分词构建词典，并按式(1)计算每个词的情感倾向，从而构建情感词典。对每个商品描述按主题进行特征分类，根据情感词典按式(2)对各特征值进行计算。其中，频数少于 10，与主题关联度高低排名超过 1000 的词不参与特征的量化，以避免小概率事件的影响。部分特征量化后的商品描述如表 2 所示。

表 2　部分商品描述的特征量化

Tab. 2　Quantification of the characteristics of some product descriptions

序号	描述文案	特征 1	特征 2	特征 3	特征 4	特征 5	特征 6
1	理由，深谙，时髦,……	23.9	16.9	30.1	18.1	26.2	25.9
2	备，点，冬日，……	10.6	9.2	19.9	15.1	9.2	5.9
3	夹克，多种，同型……	10.8	3.6	3.7	0.0	5.3	1.8
4	针织，温差，备选，……	12.7	5.5	14.7	23.2	25.2	12.6
5	轻盈，外面，羽绒,……	29.2	36.8	36.7	24.5	31.2	12.8

2.3　LightGBM 模型训练

在训练过程中，将数据集按 7：3 比例分成训练集和测试集，使用 5 折交叉验证及网格搜索(Grid Search)穷举的方式，对模型进行调参。将需要调参的参数的值分别进行训练，为以 5 折交叉验证的平均得分作为模型最优参数，然后进行一下步的调参直到调参完成，过程如图 7 所示。实线为模型在各个参数值下的 5 折交叉验证平均得分，色块上端和下端分别为得分的最高分和最低分。

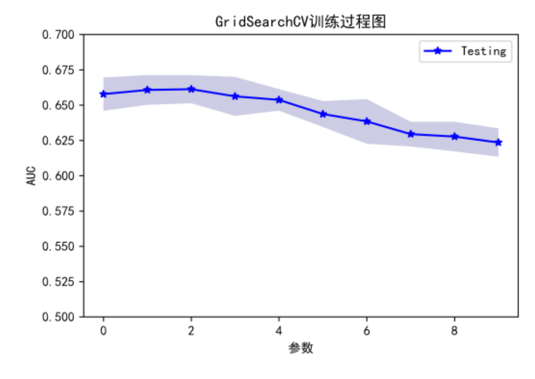


图 7　调参示例

Fig. 7　Parameter adjustment

其中，模型所训练的样本数据为不平衡数据，正负样本比值为 4.6，因此本文通过正负样本惩罚权重的方法，对分类中不同样本数量的类别分别赋予权重。即对 LightGBM 模型的参数 scale pos weight 进行设置，设置值为 5。

3　实验结果

3.1　模型的评价及对比

由于本文使用的是不平衡样本，因此本文将选用 AUC 值作为模型之间的主要评价指标。AUC 值被定义为 ROC 曲线下与坐标轴围成的面积，一般用于表示模型的综合性能，其特点是不容易受到不平衡样本的影响。以普通点击量为负例，高点击量为正例。则当样本不平衡时，若模型的预测偏向于比例大的负例时，会导致模型的准确率偏大，不能客观反映模型的性能。而对于不平衡样本，对比例小正例样本的预测识别也相当重要，召回率可以表示样本中的正例有多少被正确预测了。因此本文用准确率和召回率指标作为辅助参考指标。

另外，为了验证模型的有效性，本文将添加已有模型的对比，及与 XGBoost、随机森林、SVM、KNN 等主流分类算法进行对比。其中，“LGBM”是以主题概率量化特征构建的 LightGBM 模型，XGBoost 等算法也通过相同的调参方法(5 折交叉验证及网格搜索)进行调参，以确保对比的公平性。

模型性能对比结果如表 3 所示，从 AUC 值看，改进后的 LightGBM 模型的 AUC 值达到了 63.13%，比以主题相关性量化特征的 LightGBM 模型的 AUC 高了 3.43%，比 XGBoost、随机森林、SVM、KNN 算法分别高了 0.39%、10.02%、2.48%、8.63%，证明了式(1)能够反映消费者的情感倾向，且 LightGBM 模型性能也比其他算法更优。

表 3　模型性能对比

Tab. 3　Model performance comparison

	AUC 值	准确率	召回率	AUC 差值
改进 LGBM	63.13%	62.74%	63.73%	0.00%
LGBM	59.70%	58.53%	61.54%	3.43%
XGBoost	62.73%	62.63%	62.90%	0.39%
随机森林	53.11%	81.17%	8.81%	10.02%
SVM	60.65%	63.81%	55.65%	2.48%
KNN 分类	54.50%	77.99%	17.40%	8.63%

从准确率和召回率的角度看，随机森林、SVM、KNN 算法的准确率明显更好，但召回率极低，证明这三种算法对高点击样本的识别能力有限，模型效果差。而改进后的模型的准确率和召回率都比改进前的模型更好，再次证明了改进的有效性。

模型的学习曲线如图 8 所示，随着样本数的增加，训练集模型的得分在不断下降，而测试集的得分在不断上升，两者得分开始接近且逐渐趋于平稳。这表明随着样本数的继续增加，模型能获得更好的性能。

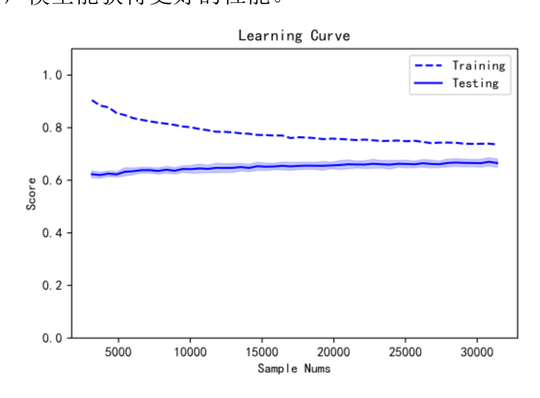


图 8　模型学习曲线

Fig. 8　Learning curve

LightGBM 算法可以通过各个特征提供的信息增益来评估特征的重要性，而特征的重要性可以作为商品各个特征对用户的整体吸引程度。特征重要性如图 9 所示，可以认为外套的“保暖”、“风格”方面的特征更能直接影响用户的点击，因此企业可以通过加大或重点宣传商品的这两方向特征，从而使生产的商品获得更高的点击。另外，企业可以根据商品不同特征的情感词典测试商品功能的组合，并通过本模型进行预测，发掘消费者需求，调整商品未来的研发方向，降低试错成本，研发更可能受到消费者青睐的产品。

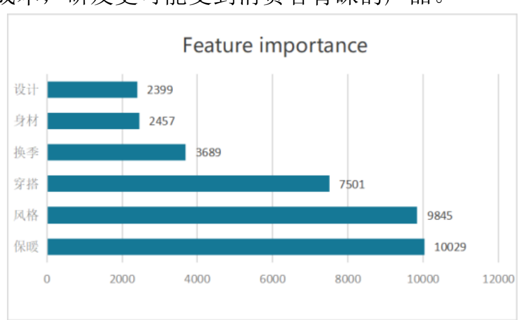


图 9　特征重要性图

Fig. 9　Feature importance

3.2　数据污染及新文本效果分析

本文中，模型的情感词典是基于所有样本构建的，存在数据污染的可能，因此本节通过分层抽样的标准 5 折交叉方法划分所有样本，训练集和验证集比例为 8：2，仅以训练集数据构建情感词典，测试集不参与情感词典的构建，从而验证新样本或新数据对模型性能的影响及模型与情感词典的关系。

如表 4 所示，5 个数据集的 AUC 值均稳定在 61-63%左右，平均 AUC 值为 62.17%，十分接近原数据集的 AUC 分数，表明模型在情感词典改变后依然有良好的性能，证明了数据污染的影响很低或没有影响，对新样本也保持着相似的预测能力。而平均值与原模型 AUC 值的差距很可能是由于构建情感词典的样本数量减少所导致的，根据大数定律，当样本足够大时，该差距无限接近于零。

表 4　新样本下的模型性能

Tab. 4　Model performance with new samples

数据集	1	2	3	4	5	平均	原数据
AUC	61.68%	62.75%	62.32%	62.55%	61.56%	62.17%	63.13%
准确率	60.60%	60.50%	61.18%	60.52%	58.94%	60.35%	62.74%
召回率	63.37%	66.27%	64.10%	65.70%	65.62%	65.01%	63.73%

3.3　冷启动问题分析

本节通过对目标商品的近邻商品的模型适应效果分析，探讨冷启动问题对模型的影响。具体做法是用相同方法筛选对比商品的商品描述文案样本，并观察其在目标商品模型中的性能表现。商品关系及模型性能如图 10 所示。

其中 AUC 是以“外套”建立的情感词典，使用各商品各自的样本训练的预测模型的性能得分，反映的是“外套”的商品特征在其他商品中的有效性；原 AUC 及原召回率则是直接使用原商品模型对其他商品样本进行预测的模型性能，反映了原商品模型对其他商品样本的预测能力；为了客观阐述商品之间的关系，除了根据商品分类划分商品外，还通过信息论中正点互信息公式(PPMI)计算总样本库中，目标商品和其他商品关键词的关系(即图 10 中各商品括号内数值)。其中 PPMI 越大，表明商品之间关联性越高。按 PMMI 排名的模型性能如表 5 所示。

对于无关商品：由图 10 和表 5 可知，“零食”与“外套”两种商品可视作无关商品。其中，“零食”的 AUC 值为 51.26%，表明“零食”仅具备“外套”很少的商品特征；原 AUC 值为 49.61%，表明原模型对“零食”商品几乎没有识别能力，与现实情况大致相符。

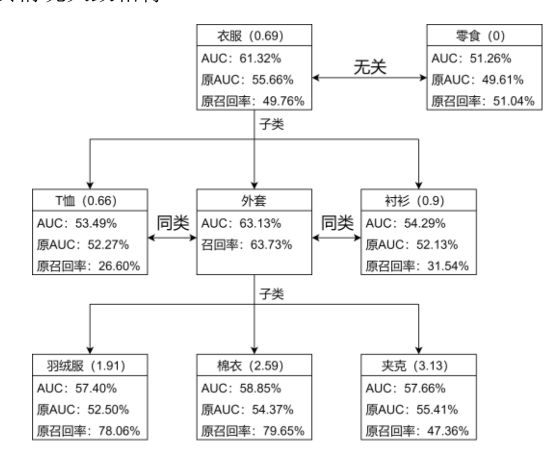


图 10　商品关系及模型性能

Fig. 10　Commodity relationships and model performance

表 5　按 PPMI 排名的商品模型性能

Tab. 5　Commodity model performance ranked by PPMI

	AUC	原模型 AUC	原模型召回率	PPMI
外套	63.13%	63.13%	63.73%	/
夹克	57.66%	55.41%	47.36%	3.13
棉衣	58.85%	54.37%	79.65%	2.59
羽绒服	57.40%	52.50%	78.06%	1.91
衬衫	54.29%	52.13%	31.54%	0.90
衣服	61.32%	55.66%	49.76%	0.69
T 恤	53.49%	52.27%	26.60%	0.66
零食	51.26%	49.61%	51.04%	0.00

对于近邻商品：由图 10 可知，模型对近邻商品的点击量依然保留着一定的识别能力。其中，直系商品(衣服、夹克、棉衣等)的 AUC 值及原 AUC 值都比同类商品(T 恤、衬衫)更高，表明用户对直系商品的特征有着更相似的偏好。从 AUC 值看，直系商品的模型性能有着明显更高的得分，而同类商品则较低，且接近无关商品的得分，这可能是由于同类商品中特征的着重点不同所导致的。与现实中，对外套、棉衣等商品的关注点与 T 恤等明显不同这一情况大致符合。从 PPMI 看，除“衣服”外，与原商品之间的关联度(PPMI)越高，则模型在该商品的适应性就越强。基于这个特性，对于缺乏历史样本的新商品，可以通过筛选与新商品的直系商品或关联度高的商品的样本进行建模，从而缓解物品的冷启动问题。企业也能通过对比近邻商品特征的情感词典，挖掘具备其他商品特征的新产品的可能。

4　结束语

本文通过挖掘商品描述文案中商品属性，构建一个基于 LightGBM 的点击预测模型。该模型可以对商品非结构化文本信息进行量化，获得用户对商品各特征的情感倾向，同时利用 LightGBM 可解释性，根据特征重要性排序识别出对商品点击影响较大的主要因素，从而可以为商品提供宣传和研发上的决策支持。针对新商品的冷启动问题，模型利用不同商品特征的相似性，使得模型能在新商品在缺少历史数据的情况下进行点击预测。实验结果证明，模型较以主题概率量化特征构建的模型具有更好的预测效果。同时，模型对新商品的预测性能与商品的关联度呈正相关。本模型从对商品描述属性对点击量的影响分析问题，用 LDA 主题模型对商品特征的划分带有一定主观性，也没有考虑到商品图片、价格等其他信息对商品点击的影响，模型性能不够高。未来工作可以通过使用更合适的主题模型及结合图像识别等技术进一步挖掘商品特征，来进一步提高模型的预测性能或可靠性。

参考文献：

[1] 张秋韵，郭斌，郝少阳，等. CrowdDepict：多源群智数据驱动的个性化商品描述生成方法 [J]. 计算机科学与探索，2020, 14 (10): 1670-1680. (Zhang Qiuyun, Guo Bin, Hao Shaoyang, *et al.* CrowdDepict: personalized recommendation content generation based on heterogeneous crowdsourced data [J]. Journal of Frontiers of Computer Science and Technology, 2020, 14 (10): 1670-1680.)

[2] Chan Zhangming, Chen Xiuying, Wang Yongliang, *et al.* Stick to the facts: Learning towards a fidelity-oriented e-commerce product description generation [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing: EMNLP-IJCNLP. 2019: 4959-4968.

[3] 刘梦娟，曾贵川，岳威，等. 面向展示广告的点击率预测模型综述 [J]. 计算机科学，2019, 46 (07): 38-49. (Liu Mengjuan, Zeng Guichuan, Yue Wei *et al.* Review on Click-through Rate Prediction Models for Display Advertising [J]. Computer Science, 2019, 46 (07): 38-49.)

[4] 朱志北，李斌，刘学军，等. 基于 LDA 的互联网广告点击率预测研究 [J]. 计算机应用研究，2016, 33 (04): 979-982. (Zhu Zhibei, Li Bin, Liu Xuejun, *et al.* Research on click-through rate prediction of Internet advertising based on LDA [J]. Application Research of Computers, 2016, 33 (04): 979-982.)

[5] Joachims T, Optimizing search engines using click through data [C]// Proceedings of the 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2002: 133-142.

[6] Guo Fan, Liu Chao, Kannan A, *et al.* Click chain model in web search [C]// Proceedings of the 18th International Conference on World Wide Web. ACM, 2009: 11-20.

[7] Graepel T, Candela J Q, Borchert T, et al, Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's bing search engine [C]// Proceedings of the 27th International Conference on Machine Learning: ICML-10. 2010: 13-20.

[8] Regelson M, Fain D. Predicting click-through rate using keyword clusters [C]// Proceedings of the Second Workshop on Sponsored Search Auctions. 2006, 9623.

[9] Richardson M, Dominowska E, Ragno R. Predicting clicks: estimating the click-through rate for new ads [C]// Proceedings of the 16th International Conference on World Wide Web. ACM, 2007: 521-530.

[10] Rendle S. Factorization machines [C]// Data Mining: ICDM, IEEE 10th International Conference on. IEEE, 2010: 995-1000.

[11] Sun Mingxuan, Lebanon G, Kidwell P, Estimating Probabilities in Recommendation Systems [J]. Applied Statistics, 2010, 61 (3) .

[12] Liu Hongyan, He Jun, Wang Tingting, et al, Combining user preferences and user opinions for accurate recommendation [J]. Electronic Commerce Research & Applications, 2013, 12 (1-6): 14-23.

[13] 于蒙，何文涛，周绪川，崔梦天，吴克奇，周文杰. 推荐系统综述 [J/OL]. 计算机应用，1-16 (2021-09-23) [2022-02-17]. http://kns. cnki. net/kcms/detail/51. 1307. tp. 20210922. 1115. 002. html, (Yu Meng, He Wentao, Zhou Xuchuan, *et al.* Review of recommendation systems [J/OL]. Journal of Computer Applications, 1-16. (2021-09-23) [2022-02-17]. http://kns. cnki. net/kcms/detail/51. 1307. tp. 20210922. 1115. 002. html)

[14] Li Xin, Xu Guandong, Chen Enhong, et al, MARS: A multi-aspect Recommender system for Point-of-Interest [C]// IEEE 31st International Conference on Data Engineering, IEEE, 2015.

[15] Fan Mingming, Khademi M, Predicting a Business Star in Yelp from Its Reviews Text Alone [J]. Computer Science, 2014.

[16] 杨贵军，徐雪，赵富强. 基于 LightGBM 算法的用户评分预测模型及应用 [J]. 数据分析与知识发现，2019, 3 (01): 118-126. (Yang Guijun, Xu Xue, Zhao Fuqiang. Predicting User Ratings with XGBoost Algorithm [J]. Data Analysis and Knowledge Discovery, 2019, 3 (01): 118-126.)

[17] 丁勇，陈夕，蒋翠清，等. 一种融合网络表示学习与 LightGBM 的评分预测模型 [J]. 数据分析与知识发现，2020, 4 (11): 52-62. (Ding Yong, Chen Xi, Jiang Cuiqing, *et al.* Predicting Online Ratings with Network Representation Learning and XGBoost [J]. Data Analysis and Knowledge Discovery, 2020, 4 (11): 52-62.)

[18] 张红丽，刘济郢，杨斯楠，等. 基于网络用户评论的评分预测模型研究 [J]. 数据分析与知识发现，2017, 1 (08): 48-58. (Zhang Hongli, Liu Jiying, Yang Sinan, *et al.* Predicting Online Users’ Ratings with Comments [J]. Data Analysis and Knowledge Discovery, 2017, 1 (08): 48-58.)

[19] 史伟，王洪伟，何绍义. 基于微博情感分析的电影票房预测研究 [J]. 华中师范大学学报：自然科学版，2015, 9 (1): 66-72. (Shi Wei, Wang Hongwei, He zhaoyi. Study on predicting movie box office based onsentiment analysis of micro-blog [J]. Journal of Central China Normal University: Natural Sciences, 2015, 9 (1): 66-72.)

[20] 孙春华，刘业政. 电影预告片在线投放对票房的影响——基于文本情感分析方法 [J]. 中国管理科学，2017, 25 (10): 151-161. (Sun Chunhua, Liu Yezheng. The Effects of Online Pre-launch Movie Trailers on the Box Office Revenue———Based on Text Sentiment Analysis Method [J]. Chinese Journal of Management Science, 2017, 25 (10): 151-161.)

[21] 李晓菊，协同过滤推荐系统中的数据稀疏性及冷启动问题研究 [D]. 华东师范大学，2018. (Li Xiaoju. Research on Data Sparsity and Cold-Start Problem in Collaborative Filtering Recommendation System [D]. East China Normal University, 2018.)

[22] Blei D M, Ng A Y, Jordan M I, Latent Dirichlet Allocation [C]// Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems, Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]. 2001.

[23] Ke G, L, Meng Qi, Finley T, *et al.* LightGBM, a Highly Efficient Gradient Boosting Decision Tree [C]// Advances in Neural Information Processing Systems 30, California, 2017, 3149-3157.

[24] Chen Qibin, Lin Junyang, Zhang Yichang, *et al.* Towards Knowledge-Based Personalized Product Description Generation in E-commerce [J]. arXiv preprint arXiv: 1903. 12457.